ConvRGX: Recognition, Generation, and Extraction for Self-trained Conversational Question Answering

Tianhua Zhang[†1,2], Liping Tang[†*1], Wei Fang[3], Hongyin Luo[3],
Xixin Wu[1,2], Helen Meng[1,2], and James Glass[3]

[1]Centre for Perceptual and Interactive Intelligence, Hong Kong SAR, China
[2]The Chinese University of Hong Kong, Hong Kong SAR, China
[3]Massachusetts Institute of Technology, Cambridge MA, USA
[1]{thzhang, lptang}@cpii.hk

Abstract

Collecting and constructing human-annotated corpora for training conversational question-answering (CQA) models has recently been shown to be inefficient and costly. To solve this problem, previous works have proposed training QA models with automatically generated QA data. In this work, we extend earlier studies on QA synthesis, and propose an efficient QA data generation algorithm under conversational settings. Our model recognizes potential dialogue topics, generates corresponding questions, and extracts answers from grounding passages. To improve the quality of generated QAs and downstream self-training of CQA models, we propose dropout and agreement-based QA selection methods. We conduct experiments on both data augmentation and domain adaptation settings. Experiments on the QuAC and Doc2Dial tasks show that the proposed method can significantly improve the quality of generated QA data, and also improves the accuracy of self-trained CQA models based on the constructed training corpora.

1 Introduction

Recent progress on pre-trained language models (Devlin et al., 2019; Clark et al., 2020; Liu et al., 2019; He et al., 2020) has significantly improved the performance of different natural language understanding tasks, including question answering (QA). However, task-specific fine-tuning of pre-trained models still requires human-annotated training corpora, especially for QA. For example, training a QA model on the Wikipedia domain needs a training set of over 80,000 human-annotated question-answer pairs (Rajpurkar et al., 2016). Annotating such a training corpus is too costly to be generalized for other domains and QA tasks. Moreover, many real-life agents answer questions in a conversational style. However, collecting data for training conversational question-answering (CQA) models is much more challenging. Recent studies have collected such corpora, but with human annotations on less than 1,000 documents (Choi et al., 2018; Feng et al., 2020).

Due to the limited amount of labeled training data and questions for conversational QA tasks being more complicated, there is a significant performance gap between single-turn and conversational QA models. As a coarse reference rather than a direct comparison, single-turn QA models achieve over 90% exact match score on SQuAD (Rajpurkar et al., 2016), and the accuracy of most CQA models is below 70% on the Doc2dial benchmark (Feng et al., 2020).

To address this problem of insufficient conversational QA for training, we propose an automatic conversational question-answering data annotation method. Inspired by the recognition-generation-extraction (RGX) pipeline (Luo et al., 2022), we design a conversation generation algorithm (named ConvRGX), which generates dialogues based on grounding documents. To generate a question and the corresponding answer in a conversation, the model first recognizes a possible dialogue topic from the grounding document, which provides information about the answer. Given the topic, a number of questions are generated. We then use a pre-trained question-answering model to verify the generated questions by comparing the selected dialogue topic and the answer extracted by the QA model given the generated questions, or the agreement between CQA models with different dropout. Among all generated QA pairs, we filter out low-quality data, samples high-quality QA pairs as the current dialogue turn, and continue to generate the next question. Compared to the baseline RGX model, we improve the answer recognition module and apply a dropout-based data selection strategy to improve the model under conversational settings.

[†]These authors contributed equally to this work and share first authorship.

[*]Now affiliated with Mohamed bin Zayed University of Artificial Intelligence. Email: liping.tang@mbzuai.ac.ae

To prove the effectiveness of ConvRGX, we evaluate the generated QA data along different dimensions. We first evaluate the question quality using Bleu (Papineni et al., 2002), RougeL (Lin, 2004), and Q-metric (Nema and Khapra, 2018). Experiments show that ConvRGX generates high-quality questions. We also conduct self-training for the CQA models with the generated QA data. Experiments show that the data generation and selection framework can constantly improve the data synthesis quality and QA self-training performance.

2 Related Work

Conversational Question Answering Recently, CQA has garnered a lot of interest, in which a QA agent answers questions from users given a piece of text as the context in a multi-turn conversation. Numerous benchmark datasets have been proposed to support investigations into different challenging facts of the CQA problem, introducing increasingly challenging aspects such as unanswerability (Reddy et al., 2019), dialogue acts (Choi et al., 2018), interpreting rules (Saeidi et al., 2018), dialogue flows (Feng et al., 2020), multiple grounding documents (Feng et al., 2021), etc. Conventional sequence models that apply various mechanisms such as attention (Choi et al., 2018) and flow (Huang et al., 2019) were explored to tackle CQA challenges. With the emergence of pre-trained language models, traditional sequence models were replaced and methods were devised to adapt these large LMs for conversations (Ohsugi et al., 2019; Qu et al., 2019). Still, challenges such as long conversational history and the lack of large training corpora exist, with various works attempting to tackle these problems (Zhao et al., 2021; Kim et al., 2021). More recently, CQA challenges have been extended with open domain retrieval (ORCQA) (Qu et al., 2020), wherein ground truth contexts are not available, which presents the need to retrive information from other sources, such as Wikipedia.

Self-trained Question Answering Recent work have studied the potential for improving QA models with question generation. A question generator benefits mutual information-based QA (Tang et al., 2017; Duan et al., 2017), in-domain data augmentation (Sachan and Xing, 2018; Puri et al., 2020; Liu et al., 2020; Klein and Nabi, 2019), and out-of-domain adaptation of QA models. Lewis et al. (2019) and Lee et al. (2020) introduced QA generation frameworks for self-trained question answering. Shakeri et al. (2020) proposed an end-to-end QA generation model, and Bartolo et al. (2021) showed that the quality and diversity of generated QA can be improved by difficult QA cases. Luo et al. (2022) proposed a cooperative self-training strategy that benefits both question generation and answering. Lewis et al. (2021); Jia et al. (2022) presented additional applications for QA generation systems.

Document-Grounded Conversation Generation In view of the prohibitive cost of manually constructing datasets, automatic conversation generation has attracted increasing research interest. One line of research (Gao et al., 2019; Gu et al., 2021) focuses on conversational question generation to produce follow-up questions based on the current dialogue context. Gao et al. (2019) generates questions with specific coreference alignment and conversation flow modeling modules, simply assuming the required answer for question generation is already predefined as input. A few efforts (Wu et al., 2022; Kim et al., 2022) attend to generate question-answering style conversations from scratch, the framework of which typically involve three components: a rationale extractor to detect the most possible text span from the grounding documents for subsequent question generation, a question generator to produce a natural question asking for information from the selected span, and an answer generator to produce answers for the questions.

3 Method

For automatically generating conversational QA data on unlabeled grounding documents, we propose a 3-step pipeline named ConvRGX. In order to generate a dialogue turn, we first recognize the upcoming possible topic from the grounding document and then generate a number of candidate questions. The generated questions are then filtered, and a pre-trained CQA model is applied to predict refined answers for the generated questions. The pipeline is illustrated in Figure 1. In this section, we introduce the details of each step of ConvRGX.

3.1 Dialogue Topic Recognition

High-quality document discourse structure are leveraged for informing dialogue flow as we synthesize question-answering conversations rather than a separate hard-to-train rationale extractor. Inspired by the findings in Gao et al. (2019) that as

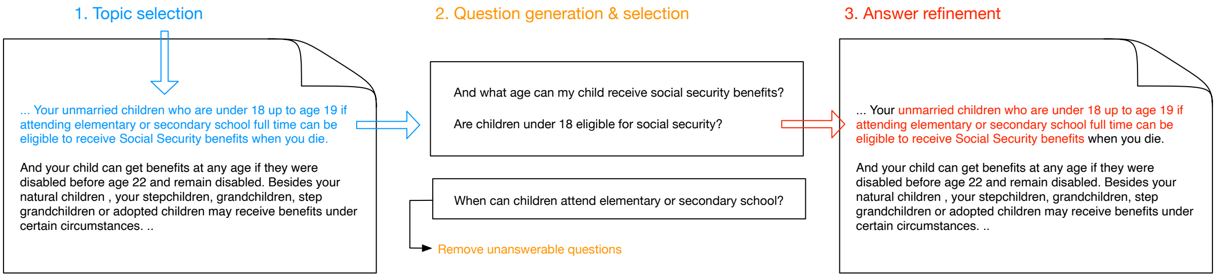
1. Topic selection
2. Question generation & selection
3. Answer refinement
... Your unmarried children who are under 18 up to age 19 if attending elementary or secondary school full time can be eligible to receive Social Security benefits when you die.
And your child can get benefits at any age if they were disabled before age 22 and remain disabled. Besides your natural children , your stepchildren, grandchildren, step grandchildren or adopted children may receive benefits under certain circumstances. ..
And what age can my child receive social security benefits?
Are children under 18 eligible for social security?
When can children attend elementary or secondary school?
Remove unanswerable questions
... Your unmarried children who are under 18 up to age 19 if attending elementary or secondary school full time can be eligible to receive Social Security benefits when you die.
And your child can get benefits at any age if they were disabled before age 22 and remain disabled. Besides your natural children , your stepchildren, grandchildren, step grandchildren or adopted children may receive benefits under certain circumstances. ..

Figure 1: The 3-step pipeline of the ConvRGX model, including dialogue topic selection, question generation & filtering, and answer refinement.

conversations progress, the focus of most questions transit from the beginning of the grounding document to the end, we design the conversation flow by heuristically following the topic flow along the document. For each document $d = \{s_1, s_2, ..., s_N\}$, we sample K instances with T ordered sentences, i.e., $I_k = \{s_{o_1}, s_{o_2}, ..., s_{o_T}\}$, where $\{o_j\}_{j=1}^{T}$ is an increasing sequence. Inspired by Dai et al. (2022), a dialogue template $C_k = \{(\triangle, s_{o_1}), (\triangle, s_{o_2}), ..., (\triangle, s_{o_T})\}$ is constructed for each instance, where $\triangle$ indicates the conversational question to be generated grounded on sentence s_{o_j} and s_{o_j} will be replaced with a refined answer a_j.

A complete conversation is generated following the determined dialogue flow in an auto-regressive manner: in the first turn, the question generator described in Section 3.2 takes as input $(\{d'\}$</s>$\{s_{o_1}\})$ with empty history $h = \varnothing$ and generates multiple diversified questions $q_1 = \{q_1^1, q_1^2, ...\}$ with text dropout at different positions. The refinement and selection module in Section 3.3 replaces s_{o_1} with a polished answer span a_1 and determines whether (q_1^i, a_1) is an answerable QA pair worth keeping. It can remove low-quality QA pairs (e.g., unanswerable QA pair) to avoid misleading the models to be trained on. The input to the question generator in the next turn is $(\{(q_1^i, a_1)\}\{d'\}$</s>$\{s_{o_2}\})$ or $(\{d'\}$</s>$\{s_{o_2}\})$, depending on whether (q_1^i, a_1) is retained or not in the previous step. The process will continue until the dialogue is complete.

To verify the effectiveness of dialogue topic flow, we also experiment with a *random* version, where the grounding passage structure is not maintained in the dialog, by shuffling the *sequential* I_k.

3.2 Conversational Question Generation

In this work, we propose a question-generation method for conversational QA. Given a document d, the related dialogue history h, and the selected dialogue topic, which is a sentence s in d, we apply an end-to-end language model to generate a question: $q = BART(\{h\}\{d'\}$ </s> $\{s\})$, where d' stands for a masked document that removes s from d. In practice, we train BART models (Lewis et al., 2020) for the question generation task with human-annotated CQA data and generate questions using the trained model with top-k sampling.

To improve the diversity in the generated questions, we propose a text dropout strategy. We randomly replace up to $len(grounding\,sentence)/10$ tokens for each selected grounding sentence with <mask>. This method is applied in the training stage of the question generator, while no text dropout is applied in the evaluation stage to maintain as much information as possible for question synthesis.

3.3 Data Selection and Answer Refinement

To ensure the quality of generated questions, we train a CQA model on human-annotated data and then utilize it to perform data selection. Specifically, given the generated question q_i^j and the dialogue history h, the pre-trained CQA model is used to extract answers from the grounding passage P. Based on the extracted answers, we propose the following two data selection strategies[1]:

Overlap-based Data Selection was previously used by Wu et al. (2022) and Kim et al. (2022). Under this strategy, we keep all questions that produce answers that overlap with QG-grounding sentences.

Dropout-based Data Selection is inspired by test-time dropout (Kamath et al., 2020), which ensembles the prediction across multiple dropout masks to deal with out-of-domain data. During the inference stage, we enable the standard dropout

[1] CQA extracted answers instead of the QG-selected sentences are kept in the generated data.

(Srivastava et al., 2014) in Transformers (Vaswani et al., 2017). We use the pre-trained CQA model to extract M answers from M different dropout masks given the same input and keep the generated question if there are at least C consistent answers among the M extracted answers.[2]

Answer Refinement After selecting high-quality questions, we propose to refine the answer using the aforementioned pre-trained conversational QA model. Specifically, after obtaining a consistent answer from the dropout-based data selection step or an answer overlapping with the QG-grounding sentence, we extend the answer to compact sentences and add one sentence before and one sentence after the answer sentences given the entire passage. The motivation is that it may be easier for the conversational QA model to predict a more accurate answer from a shorter span than from a long passage.

3.4 QA Self-training

After selecting high-quality QA data, we construct synthetic training corpora for fine-tuning the pre-trained QA models. In this work, we train extractive question-answering models with the RoBERTa (Liu et al., 2019) backbone. A linear model is stacked on the pre-trained RoBERTa model to predict the starting and ending positions of the answers as standard extractive QA baselines.

4 Experimental Setup

4.1 Datasets

Doc2Dial (Feng et al., 2020) contains goal-oriented dialogues that are grounded in documents. It consists of two subtasks: 1) grounding span identification based on dialogue context; and 2) agent response generation based on extracted spans. In this paper, we focus on the first subtask and evaluate the performance of ConvRGX under the machine reading comprehension setting. During training and auto-regressive dialogue generation, we replace the agent response in dialogue history by its *oracle* and *identified* grounding spans respectively since no agent response generation module is involved in ConvRGX.

Question Answering in Context (QuAC) (Choi et al., 2018) is a standard CQA benchmark that contains questions that are complex, highly context-dependent and sometimes unanswerable. Dialogue answers in this dataset, if available, are spans within the given grounding context. We evaluate our approach with the standard split. Dataset statistics for the two experimental settings described in Section 4.3 are shown in Appendix A.

4.2 Question Generation Evaluation

We first evaluate the quality of questions generated by ConvRGX on the two CQA datasets. Since there has always been criticism for evaluating the performance of automatic generation by common n-gram-based similarity metrics only, Q-metric was proposed to measure the answerability of generated question-answer pairs. It scores the quality of a generated question by assigning different importance to four types of information: named entity, question type, relevant content and function words, which correlates better with human judgment. We refer interested readers to Nema and Khapra (2018) for more details. We report the question generation performance using 1) traditional automatic evaluation metrics: Bleu (Papineni et al., 2002) and RougeL (Lin, 2004); and 2) their corresponding Q-Metrics: Q-Bleu and Q-RougeL.

4.3 Question Answering Evaluation

To show the effectiveness of ConvRGX on addressing the data scarcity issue for conversational QA tasks, we evaluate ConvRGX on downstream CQA tasks under two settings: 1) extractive-based CQA, and 2) retrieval-based CQA.

Extractive-based CQA The extractive-based CQA follows the general evaluation step of QA systems. Specifically, we concatenate the conversational question, i.e., the concatenation of the entire dialogue history and the current question, with the grounding passage as input to a Roberta-large model. We then stack a linear model on the Roberta-large model to predict the starting and ending positions of the answers. Under this setting, we follow Feng et al. (2020) to evaluate ConvRGX with Exact match (EM) and F1 score for Doc2Dial. Following Choi et al. (2018), we report results with F1 score and human equivalence score (HEQ) for QuAC.

Retrieval-based CQA Besides the commonly used extractive-based CQA setting, we further consider the retrieval-based CQA tasks. Specifically, given a new conversational question, we search via BM25[3] (Robertson et al., 2009) from the database

[2] In our setting, M is set as 5 and C is set as 4.

[3] In our preliminary experiments, SimCSE (Gao et al., 2021) was also used for training the representations of con-

that consists of all conversational queries in the dataset (original training dataset or the dataset augmented by synthetic data). Then we treat the answer of the retrieved conversational query as the final answer. To be consistent with the EM and F1 values in extractive CQA, we measure the retrieval-based CQA with EM@k (EM@1, EM@5, EM@10) and F1@k (F1@1, F1@5, F1@10), i.e., the highest EM and F1 values among top k retrieved answers.

For both extractive and retrieval CQA, we perform experiments under the following two settings:

Data Augmentation Under this setting, we use the original dataset splits that are given and train on the training set. During evaluation, ConvRGX generates new QA pairs using documents from both the training and validation sets, which are then used to augment the original training set. The augmented dataset is finally used to train the QA model.

Unseen Documents We also evaluate our approach on a set of unseen documents to investigate the generalization performance. To prevent leakage, we remove the training dialogues that are based on documents in the validation set for Doc2Dial. For QuAC, the documents do not overlap between training and validation, so the original splits already correspond to this setting.

Implementation details for both QG and QA settings can be found in Appendix B.

4.4 Baselines

Since there is no prior performance benchmark that is readily available[4], we compare the proposed model against three baselines typically used for natural language data augmentation, as done by Wu et al. (2022).

Easy Data Augmentation (EDA) EDA is proposed by Wei and Zou (2019) to augment data through text editing operations. In particular, EDA consists of four simple operations: synonym replacement, random insertion, random swap, and random deletion.

Back-translation Back-translation augments natural language data by first translating the text into a second language and then back-translating them to the original language. Following BERT-QA (Chadha and Sood, 2019) and DG2 (Wu et al., 2022), we translate all user utterances to French and then translate them back to English.[5]

Paraphrase Paraphrasing rewrites text using different words or sentence structures. In particular, we use the BART-large model (Lewis et al., 2020) trained on the MRPC (Dolan and Brockett, 2005), QQP (Shankar et al., 2017) and PAWS (Zhang et al., 2019) datasets.[6]

5 Results

5.1 Question Generation Quality

We conduct intrinsic evaluation of question generation when text dropout is (*w/*) and is not (*w/o*) introduced during the training and validation process with Q-metric and n-gram-based metrics. As shown in Table 1(a), the text dropout strategy in question generator during training achieves significant performance improvement at all metrics when evaluated on the Doc2Dial validation set under both *w/* and *w/o* dropout settings. On QuAC in Table 1(b), text dropout training obtains performance comparable to the baseline when evaluated *w/o* dropout, and improves the performance when evaluated *w/* dropout. Training with explicit dropout introduces variations to the input and enhances the robustness of the question generator.

Comparing *w/* and *w/o* dropout during the validation process, text dropout on the validation input decreases the performance since keywords in the grounding span might be masked, resulting in information loss and failure to derive the ground-truth question. On the other hand, text dropout during auto-regressive dialogue generation can boost the downstream CQA performance. Specifically, by masking different input positions, questions focusing on different information will be generated, which enriches the diversity of synthesized dialogues. It reveals the discrepancy between the intrinsic question generator performance measurements and the actual need for high-quality and diverse conversations to increase CQA performance, showing the necessity of evaluating data quality by extrinsic downstream CQA performance. We report the corresponding results in Section 6.2.

versational queries, which gave us similar results but required more training time than BM25.

[4] The two works that we have found to be closest to our settings are Wu et al. (2022) and Kim et al. (2022). However, the codes of these two works are not publicly available.

[5] Translated via the Google Translate API.

[6] https://huggingface.co/eugenesiow/bart-paraphrase

					(a) Doc2Dial						
Train	**Val**	**Q-Bleu1**	**Q-Bleu2**	**Q-Bleu3**	**Q-Bleu4**	**Q-RougeL**	**Bleu1**	**Bleu2**	**Bleu3**	**Bleu4**	**RougeL**
w/o	w/o	0.321	0.28	0.262	0.252	0.333	0.272	0.153	0.099	0.068	0.309
w/	w/o	**0.327**	**0.287**	**0.268**	**0.257**	**0.339**	**0.282**	**0.164**	**0.108**	**0.078**	**0.318**
w/o	w/	0.318	0.279	0.260	0.250	0.332	0.269	0.152	0.097	0.067	0.308
w/	w/	**0.323**	**0.283**	**0.265**	**0.254**	**0.336**	**0.279**	**0.162**	**0.106**	**0.076**	**0.316**
					(b) QuAC						
w/o	w/o	0.332	0.296	**0.274**	**0.263**	0.341	0.288	**0.182**	**0.116**	**0.085**	0.315
w/	w/o	**0.333**	0.296	0.273	0.262	**0.342**	**0.289**	0.180	0.112	0.080	**0.316**
w/o	w/	0.325	0.289	0.267	0.257	0.335	0.279	0.174	0.109	0.079	0.306
w/	w/	**0.329**	**0.292**	**0.270**	**0.259**	**0.339**	**0.285**	**0.177**	**0.110**	**0.079**	**0.313**

Table 1: Question generator performance on the Doc2Dial and QuAC validation sets. *w/* and *w/o* indicate whether text dropout is introduced during training and validation process of the question generator.

Training Data		*Data Augmentation*				*Unseen Documents*			
Original	**Generated**	**best EM**	**cor F1**	**cor EM**	**best F1**	**best EM**	**cor F1**	**cor EM**	**best F1**
✓	✗	62.93	75.78	62.87	75.90	49.07	67.43	48.46	67.86
✓	EDA	64.07	75.35	63.29	75.88	50.07	67.70	49.66	68.16
✓	Back-translation	63.92	75.82	62.98	75.91	49.66	67.95	49.18	68.17
✓	Paraphrase	63.98	75.89	63.82	76.07	49.37	68.01	49.25	68.15
✓	ConvRGX	**64.62**	**77.05**	**64.39**	**77.31**	**50.91**	**69.26**	**50.80**	**69.34**
✗	ConvRGX	50.88	67.84	50.80	67.87	43.67	63.00	43.28	63.06

Table 2: Extractive-based question answering performance on the Doc2Dial validation set. We report the best EM together with the corresponding F1 scores and the best F1 together with the corresponding EM scores.

Training Data		*Unssen Documents*		
Original	**Generated**	**best F1**	**HEQ-Q**	**HEQ-D**
✓	✗	71.12	67.46	11.17
✓	EDA	70.46	66.88	10.80
✓	Back-translation	70.67	66.88	9.70
✓	Paraphrase	70.47	66.54	10.00
✓	ConvRGX	**71.64**	**68.37**	**12.70**
✗	ConvRGX	55.58	49.57	3.80

Table 3: Extractive-based question answering performance on the QuAC validation set.

5.2 Self-trained CQA Results

In this section, we validate the data generation quality of ConvRGX by training QA models on the generated data and compare the self-training performance. We assess the accuracy of extractive and retrieval-based QA models.

5.2.1 Extractive-based CQA

Doc2dial The experiments on the Doc2dial benchmark is shown in Table 2. We report the best EM together with the corresponding F1 scores and the best F1 together with the corresponding EM scores. We first show the in-domain self-training results, where ConvRGX generates synthetic data on seen documents. The experiment results show that most data augmentation models outperforms the model trained only with the human-generated training set, validating our hypothesis that augmenting the training corpus can benefit CQA models. On the other hand, ConvRGX outperforms all data augmentation models across different metrics. Among all baseline models, EDA achieves the best EM score but the corresponding F1 is worse than the unaugmented baseline. On the other hand, the paraphrase method achieves the best F1. We found that the F1 improvement of ConvRGX over paraphrasing is higher than the performance gap between paraphrasing and the base CQA model, indicating that the improvement of the ConvRGX model is more significant.

To test the generalization ability of this approach, we train the models on a subset of documents and generate data on unseen documents. The results shown in Table 2 indicate that the ConvRGX model achieves more improvement than the in-domain setting. This validates our hypothesis that CQA performance can benefit from the QA data synthesis approach on unlabeled documents.

QuAC We evaluate the model performance on the QuAC and present the experiment results in Table 3, where all test documents are unseen in the training process. The ConvRGX model still outperforms all baselines, but the improvement is not as high as the Doc2dial model, because the

Training Data		*Data Augmentation*						*Unseen Documents*					
Original	**Generated**	**EM@1**	**EM@5**	**EM@10**	**F1@1**	**F1@5**	**F1@10**	**EM@1**	**EM@5**	**EM@10**	**F1@1**	**F1@5**	**F1@10**
✓	✗	11.86	26.18	33.06	23.30	43.13	51.05	0.86	2.14	2.82	12.07	22.75	26.75
✓	EDA	11.66	22.08	28.73	23.06	38.15	46.12	0.81	1.74	2.42	11.95	20.27	24.27
✓	Back-translation	11.98	21.90	28.05	23.38	37.98	45.44	0.88	1.74	2.24	12.08	20.19	24.01
✓	Paraphrase	11.98	22.51	28.15	23.40	38.55	45.62	0.83	1.76	2.32	12.13	20.41	24.03
✓	ConvRGX	**13.17**	**30.72**	**39.35**	**26.22**	**48.02**	**56.85**	**6.34**	**12.71**	**17.15**	**20.59**	**34.43**	**40.92**

Table 4: Retrieval-based question answering performance on the Doc2Dial validation set.

number of annotated QuAC QA samples is much larger than Doc2dial, which reduces the potential of data augmentation. The result further validates our conclusion that the performance of CQA models is significantly affected by the data annotation effort.

5.2.2 Retrieval-based CQA

We report the retrieval-based CQA performance on Doc2Dial dataset in Table 4. ConvRGX outperforms all baselines across all evaluation metrics under both data augmentation and unseen documents settings for retrieval-based CQA tasks. Different from the results of extractive CQA setting, adding more data generated from the three baselines leads to a performance reduction on EM@k and F1@k (k=5,10) under the retrieval-based CQA setting. The reason is that the three data augmentation baselines cannot improve the answer text coverage of conversational questions in the dataset and can only generate semantically similar questions.

Under the unseen documents setting, the EM values of data augmentation baselines and the baseline of no data augmentation are almost zero. This is because the grounding documents in the testing dataset are invisible in the training dataset. Thus almost no conversational questions are grounded on the answer texts in the testing dataset.[7] After data augmentation via the three baselines, no new answer text is involved. On the contrary, ConvRGX obtains higher EM and F1 values because it can generate conversational questions on the documents of the testing dataset to cover more possible answer texts. More experimental results on QuAC are in Appendix D.

6 Analysis

In this section, we report results with extractive-based CQA models trained with only generated dialogues on the Doc2Dial dataset to further explore the contribution of different factors to the quality of synthesized conversations.

(a) No Data Selection

Dial-Num	**Selection**	**Flow-Order**	**EM**	**F1**
single	-	random	**2.69**	**31.31**
		sequential	2.44	22.39

(b) Best Setting without Answer Refinement

Setting	**Flow-Order**	**EM**	**F1**
Data Augmentation	random	50.71	66.70
	sequential	**51.49**	**66.81**
Unseen Documents	random	43.03	62.99
	sequential	**43.67**	**63.00**

Table 5: Ablation study of different topic recognition strategies (*random* versus *sequential*). Best setting without answer refinement refers to QG Dropout, *dropout* selection and *multi* Dial-Num.

6.1 Effect of Topic Recognition Strategies

Since the synthesized dialogue flows are constructed from high-quality documents, we analyze the effect of topic recognition from two aspects.

Document Sentence Sampling (*Dial-Num*) We examine the performance of using all sentences in a short truncated passage to derive a single dialogue template (*single*), i.e., $K = 1, T = N$ and sampling different sets of sentences to construct multiple dialogue templates from a enlarged passage (*multi*), i.e., $K > 1, T < N$. Intuitively, the *multi* setting considers different possible combinations of sentences and hence increases the information-coverage and diversity of subsequent dialogue generation. The EM and F1 scores are raised from 47.03 to 50.71 and 62.56 to 66.70 respectively when the setting is changed from *single* to *multi*, in line with our expectations. More implementation details are in Appendix C.

Document Sentence Order (*Flow-Order*) We further analyze how the sentence usage order affect the data quality. The *sequential* approach derives dialogue topic flow following the document discourse and aims to generate dialogues conversing each topic (possibly in depth) before shifting to another one. On the contrary, the *random* approach increases the variability to the possible dialogues

[7] The EM values are close to zero but not exactly zero because there are some sentences overlapping among the documents in the training dataset and testing dataset even under the unseen documents setting.

Selection	Dial-Num	QG Dropout	EM	F1
overlap	single	✗	35.96	54.44
		✓	**41.99**	**59.99**
overlap	multi	✗	41.21	60.00
		✓	**44.24**	**62.56**
dropout	single	✗	42.21	59.01
		✓	**47.03**	**62.56**
dropout	multi	✗	49.10	65.37
		✓	**50.71**	**66.70**

Table 6: Ablation study of QG Text Dropout (✗ versus ✓) and Data Selection strategies (*overlap* versus *dropout*). This table reports the best EM and corresponding F1 scores with *random* flow. (Remark: if *sequential* flow were used, similar results are observed.)

Flow-Order	Answer Refinement	EM	F1
random	✗	35.96	54.44
	✓	**40.05**	**58.07**
sequential	✗	36.71	54.15
	✓	**38.61**	**56.18**

Table 7: Ablation study of answer refinement (✗ versus ✓). We use *overlap* selection and *single* Dial-Num, *without* QG dropout.

and is closer to information-asymmetric situation, where no pre-knowledge of the document is given to questioners. We found that there is no single conclusion that one setting is superior to the other in all cases since the performance depends on the overall design. In particular, *sequential* relies heavily on the data selection strategy. Without turn-level filtering, the *single-sequential* strategy produces conversations strictly following the grounding document structure, which hurts the QA self-training since the QA model can simply predict next sentence in the document as the answer. Experimental results in Table 5(a) verifies this hypothesis. In our best setting for *Data Augmentation* and *Unseen Documents*, *sequential* results exceed *random* setting with well-designed selection strategy eliminating low-quality QA pairs and introducing variations to resolve the aforementioned limitation.

6.2 Effect of Text Dropout

Table 6 presents the extrinsic evaluation results of the synthesized conversations when text dropout is (✓) and is not (✗) introduced during the question generation process. We measure the quality of generated dialogues using the extractive-based QA performance. Introducing text dropout in question generation gave marked improvement under all four settings. In particular, 16.8% improvement is observed when QG text dropout is involved in the *overlap-selection* and *single-Dial-Num* settings. By masking different positions of the input grounding span, ConvRGX generates not only diverse questions of various forms with similar semantic meanings, but also information-seeking questions for different knowledge. An example is shown in Appendix E. The contrasting results to Section 5.1 imply that evaluating the generated question quality by intrinsic Q-metric is not sufficient.

6.3 Effect of Data Selection and Anwer Refinement

Table 6 also shows how the data selection in conversation generation contributes to the improvement of CQA performance. First, the QA model performs poorly on both EM and F1 when trained on generated conversations without data selection. Introducing data selection boosts the performance drastically, with the EM score increasing from 2.69 to 42.21 (35.96) and F1 score from 31.31 to 59.01 (54.44) if dropout(overlap)-based selection is involved. Although the question generator can produce diverse questions of fluency and coherence, the generated questions may not be answerable by the given grounding text s. We design selection strategies that replace the grounding span with a fine-grained answer a and filter out low-quality, especially unanswerable QA pairs. An example is shown in Appendix E. We also find that the dropout-based selection strategy outperforms the overlap-based strategy significantly in all settings listed in Table 6. It implies that the data selection strategies has a significant effect on the CQA generation quality.

Table 7 shows how answer refinement affects the performance of CQA. As shown in Table 7, adding answer refinement improve the performance of CQA, with the EM score increasing from 35.96 (36.71) to 40.05 (38.61) and F1 score from 54.44 (54.15) to 58.07 (56.18) with the *random* (*sequential*) flow-order, which implies the positive effect of answer refinement in improving generation quality.

7 Conclusion

We propose ConvRGX, an automatic CQA data annotation method extended from the recognition-generation-extraction (RGX) framework for conversational applications, which can generate high-quality CQA data that can be used for question generation and data augmentation. We demonstrate the effectiveness of ConvRGX on standard conversational benchmarks, which show improvements over current data generation and augmentation methods for both question quality and self-training performance. In summary, ConvRGX presents a scalable and effective way to approach CQA problems that have limited human annotation.

Acknowledgement

This research was supported by the Center for Perceptual and Interactive Intelligence (CPII) Ltd under the Innovation and Technology Commission's InnoHK Scheme.

Limitations

This paper proposes a conversational QA generation system and evaluates the generated QA quality on two publicly available benchmarks. Although the improved Q-metric indicates that the ConvRGX generates better QA data than previous methods, the QA generation pipeline can still generate noisy conversations and hence cannot be entirely trusted. On the other hand, if we force the model such that only QA pairs with high confidence are selected, the diversity of generated data would be limited. In the future, we will investigate the trafeoff between reliability and diversity in CQA generation tasks.

Ethics Statement

The CQA generation system proposed in this work can augment the performance of CQA models, but also introduce the following risks. Firstly, the questions are generated according to grounding documents. As a result, they might deliver social biases and misinformation contained in the documents. Secondly, the method increases the size of CQA corpora and the computational cost of model training. Lastly, since the system can automatically annotate unlabeled documents, it might reduce the number of jobs in manual data annotation.

References

Max Bartolo, Tristan Thrush, Robin Jia, Sebastian Riedel, Pontus Stenetorp, and Douwe Kiela. 2021. Improving question answering model robustness with synthetic adversarial data generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8830–8848, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ankit Chadha and Rewa Sood. 2019. BERTQA - attention on steroids. *CoRR*, abs/1912.10435.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2174–2184, Brussels, Belgium. Association for Computational Linguistics.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: pre-training text encoders as discriminators rather than generators. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Zhuyun Dai, Arun Tejasvi Chaganty, Vincent Zhao, Aida Amini, Qazi Mamunur Rashid, Mike Green, and Kelvin Guu. 2022. Dialog inpainting: Turning documents into dialogs.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

William B. Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

Nan Duan, Duyu Tang, Peng Chen, and Ming Zhou. 2017. Question generation for question answering. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 866–874, Copenhagen, Denmark. Association for Computational Linguistics.

Song Feng, Siva Sankalp Patel, Hui Wan, and Sachindra Joshi. 2021. MultiDoc2Dial: Modeling dialogues grounded in multiple documents. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6162–6176, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Song Feng, Hui Wan, Chulaka Gunasekara, Siva Patel, Sachindra Joshi, and Luis Lastras. 2020. doc2dial: A goal-oriented document-grounded dialogue dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8118–8128, Online. Association for Computational Linguistics.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, *EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 6894–6910. Association for Computational Linguistics.

Yifan Gao, Piji Li, Irwin King, and Michael R. Lyu. 2019. Interconnected question generation with coreference alignment and conversation flow modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4853–4862, Florence, Italy. Association for Computational Linguistics.

Jing Gu, Mostafa Mirshekari, Zhou Yu, and Aaron Sisto. 2021. ChainCQG: Flow-aware conversational question generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2061–2070, Online. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Hsin-Yuan Huang, Eunsol Choi, and Wen-tau Yih. 2019. Flowqa: Grasping flow in history for conversational machine comprehension. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Robin Jia, Mike Lewis, and Luke Zettlemoyer. 2022. Question answering infused pre-training of general-purpose contextualized representations. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 711–728, Dublin, Ireland. Association for Computational Linguistics.

Amita Kamath, Robin Jia, and Percy Liang. 2020. Selective question answering under domain shift. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5684–5696, Online. Association for Computational Linguistics.

Gangwoo Kim, Hyunjae Kim, Jungsoo Park, and Jaewoo Kang. 2021. Learn to resolve conversational dependency: A consistency training framework for conversational question answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6130–6141, Online. Association for Computational Linguistics.

Gangwoo Kim, Sungdong Kim, Kang Min Yoo, and Jaewoo Kang. 2022. Towards more realistic generation of information-seeking conversations. *CoRR*, abs/2205.12609.

Tassilo Klein and Moin Nabi. 2019. Learning to answer by learning to ask: Getting the best of gpt-2 and bert worlds. *arXiv preprint arXiv:1911.02365*.

Dong Bok Lee, Seanie Lee, Woo Tae Jeong, Donghwan Kim, and Sung Ju Hwang. 2020. Generating diverse and consistent QA pairs from contexts with information-maximizing hierarchical conditional VAEs. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 208–224, Online. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Patrick Lewis, Ludovic Denoyer, and Sebastian Riedel. 2019. Unsupervised question answering by cloze translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4896–4910, Florence, Italy. Association for Computational Linguistics.

Patrick Lewis, Yuxiang Wu, Linqing Liu, Pasquale Minervini, Heinrich Küttler, Aleksandra Piktus, Pontus Stenetorp, and Sebastian Riedel. 2021. PAQ: 65 million probably-asked questions and what you can do with them. *Transactions of the Association for Computational Linguistics*, 9:1098–1115.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Dayiheng Liu, Yeyun Gong, Jie Fu, Yu Yan, Jiusheng Chen, Jiancheng Lv, Nan Duan, and Ming Zhou. 2020. Tell me how to ask again: Question data augmentation with controllable rewriting in continuous space. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5798–5810, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Hongyin Luo, Shang-Wen Li, Mingye Gao, Seunghak Yu, and James Glass. 2022. Cooperative self-training of machine reading comprehension. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 244–257, Seattle, United States. Association for Computational Linguistics.

Preksha Nema and Mitesh M. Khapra. 2018. Towards a better metric for evaluating question generation systems. In *Proceedings of the 2018 Conference on*

Empirical Methods in Natural Language Processing, pages 3950–3959, Brussels, Belgium. Association for Computational Linguistics.

Yasuhito Ohsugi, Itsumi Saito, Kyosuke Nishida, Hisako Asano, and Junji Tomita. 2019. A simple but effective method to incorporate multi-turn context with BERT for conversational machine comprehension. In *Proceedings of the First Workshop on NLP for Conversational AI*, pages 11–17, Florence, Italy. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Raul Puri, Ryan Spring, Mohammad Shoeybi, Mostofa Patwary, and Bryan Catanzaro. 2020. Training question answering models from synthetic data. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5811–5826, Online. Association for Computational Linguistics.

Chen Qu, Liu Yang, Cen Chen, Minghui Qiu, W Bruce Croft, and Mohit Iyyer. 2020. Open-retrieval conversational question answering. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 539–548.

Chen Qu, Liu Yang, Minghui Qiu, Yongfeng Zhang, Cen Chen, W Bruce Croft, and Mohit Iyyer. 2019. Attentive history selection for conversational question answering. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 1391–1400.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. CoQA: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Mrinmaya Sachan and Eric Xing. 2018. Self-training for jointly learning to ask and answer questions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 629–640, New Orleans, Louisiana. Association for Computational Linguistics.

Marzieh Saeidi, Max Bartolo, Patrick Lewis, Sameer Singh, Tim Rocktäschel, Mike Sheldon, Guillaume Bouchard, and Sebastian Riedel. 2018. Interpretation of natural language rules in conversational machine reading. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2087–2097, Brussels, Belgium. Association for Computational Linguistics.

Siamak Shakeri, Cicero Nogueira dos Santos, Henghui Zhu, Patrick Ng, Feng Nan, Zhiguo Wang, Ramesh Nallapati, and Bing Xiang. 2020. End-to-end synthetic data generation for domain adaptation of question answering systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5445–5460, Online. Association for Computational Linguistics.

Iyer Shankar, Dandekar Nikhil, and Csernai Kornel. 2017. First quora dataset release: question pairs (2017). *URL https://www. quora. com/q/quoradata/First-Quora-Dataset-Release-Question-Pairs*.

Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1):1929–1958.

Duyu Tang, Nan Duan, Tao Qin, Zhao Yan, and Ming Zhou. 2017. Question answering and question generation as dual tasks. *arXiv preprint arXiv:1706.02027*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Jason W. Wei and Kai Zou. 2019. EDA: easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 6381–6387. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771.

Qingyang Wu, Song Feng, Derek Chen, Sachindra Joshi, Luis A. Lastras, and Zhou Yu. 2022. DG2: data augmentation through document grounded dialogue generation. In *Proceedings of the 23rd Annual Meeting* *of the Special Interest Group on Discourse and Dialogue, SIGDIAL 2022, Edinburgh, UK, 07-09 September 2022*, pages 204–216. Association for Computational Linguistics.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. PAWS: Paraphrase adversaries from word scrambling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1298–1308, Minneapolis, Minnesota. Association for Computational Linguistics.

Jing Zhao, Junwei Bao, Yifan Wang, Yongwei Zhou, Youzheng Wu, Xiaodong He, and Bowen Zhou. 2021. RoR: Read-over-read for long document machine reading comprehension. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1862–1872, Punta Cana, Dominican Republic. Association for Computational Linguistics.

A Dataset Statistics

Table 8 shows the statistics of the Doc2dial and QuAC benchmarks.

B Implementation Details

All model training and inference in this work are conducted on a single NVIDIA RTX A6000 (48G).

CQG We implement the question generator using Bart-large based on the Transformers library (Wolf et al., 2019). For QG text dropout, we randomly replace $max(1, len(grounding\,span)/10)$ tokens with `<mask>`. For Doc2Dial, the number of training epochs is set as 8 with a batch size of 4 and evaluation is conducted after each epoch. We use a learning rate of 3e-5 with a weight decay of 0.001. Maximum input sequence length is set as 1024 with dialogue history length no longer than 128 due to the long document. Maximum target length length is set as 128. During inference time, we initiate multi-generation. For each grounding sentence, we generate 5 questions. If text dropout is introduced during dialogue generation, we randomly mask the grounding sentence 5 times to introduce different variations to the input.

CQA Roberta-large based on the Transformers library (Wolf et al., 2019) is used as the model backbone of CQA tasks. The number of training epochs is set as 5 and the number of evaluation steps is set as 2000 with a batch size of 15. We use a learning rate of 3e-5 with a weight decay of 0.01. Maximal sequence length is set as 512 and maximal answer length is set as 50. The number of tokens for document stride is set as 128 and the number of warmup steps is set as 1000. The whole training process takes about 3.5 hours for Doc2Dial and 8 hours for QuAC without any synthetic data.

C Details of Document Sentence Sampling

Table 9 shows the data statistics of the passage truncation and dialogue template construction for Doc2Dial datasets reported in this paper.

Single Dial-Num We truncate each Doc2Dial document into passages with $N = 6$ sentences and obtain 6979 unique passages in total. Then following the ConvRGX generation pipeline, each document sentence is regarded as the grounding span to generate a QA turn in an auto-regressive manner. Hence, we produce a single dialogue template with 6 turns (i.e., QA pairs) per passage.

Multi Dial-Num We first enlarge each passage to $N = 12$ sentences to enable multi-template construction. For each truncated passage, we sample at most $K = 3$ sets of sentences, with $T = 8$ sentences for each set and let any two sets have 4 different sentences. Overall, we get 2794 truncated passages and 7162 dialogue templates. To verify the contribution of longer passage truncation and longer dialogue turn to the performance improvement of the extractive-based QA model, we also implement the setting of $K = 3$ and $T = 6$ as a comparison to the *single* setting. Experimental results indicate that both longer passage and longer dialogue turn can bring benefits to the quality of synthesized conversations since the *multi* setting is closer to the statistics of human annotated Doc2Dial dataset.

D Retrieval-based CQA Results on QuAC

Table 10 shows the retrieval-based CQA results on QuAC dataset.

E Qualitative Results

Table 11(a) shows an example where ConvRGX generates multiple information-seeking questions with different text dropout. Table 11(b) shows an example that our question generator produces a grammatically correct but unanswerable question by the passage with the text in italics as grounding span. Our data selection module succeeds in identifying and eliminating such kinds of instances.

Doc2Dial	*Train*				*Val*				*# doc overlap*
	# dial	**# doc**	**# tok/doc**	**# tok/usr**	**# dial**	**# doc**	**# tok/doc**	**# tok/usr**	
Data Augmentation	3474	402	833	10.2	661	272	821	10.0	237
Unseen Documents	1345	166	899	10.4					0
QuAC	*Train*				*Val*				*# doc overlap*
	# dial	**# doc**	**# tok/doc**	**# tok/usr**	**# dial**	**# doc**	**# tok/doc**	**# tok/usr**	
Unseen Documents	11567	6843	396.8	6.5	1000	1000	440	6.5	0

Table 8: Data statistics of Doc2Dial and QuAC datasets used in our experiments. The number of documents are obtained after document deduplication. Models are trained on the *Train* set and evaluated on the *Val* set.

	# s/psg	# turn	# dial/psg	# diff s	# psg	# template
single	6	6	1	/	6979	6979
multi	12	8	3	4	2794	7162

Table 9: Data statistics of passage truncation and dialogue template construction on Doc2Dial dataset.

Training Data		*Unseen Documents*		
Original	**Generated**	**F1@1**	**F1@5**	**F1@10**
✓	✗	2.98	8.13	10.39
✓	EDA	2.98	6.89	8.93
✓	Back-translation	2.89	7.06	9.14
✓	Paraphrase	3.04	7.09	9.15
✓	ConvRGX	**4.69**	**9.06**	**11.02**

Table 10: Retrieval-based question answering performance on the QuAC validation set.

Table 12 demonstrates a complete dialogue generated by ConvRGX grounded on the given passage in an auto-regressive manner.

(a) Diverse question generation example
[Passage]: ... *Your unmarried children who are under 18 up to age 19 if attending elementary or secondary school full time can be eligible to receive Social Security benefits when you die.* And your child can get benefits at any age if they were disabled before age 22 and remain disabled. Besides your natural children , your stepchildren, grandchildren, step grandchildren or adopted children may receive benefits under certain circumstances. ...
[Generated questions]: And what about my unmarried children, who are under 18 years old? And at what age can my child receive Social Security benefits? And what about my children who are not in school? Are children under 18 eligible for Social Security?
(b) Unswerable question filtering example
[Passage]: Your loan servicer can help you understand your options. You may be able to switch repayment plans to get a lower monthly payment , consider an income - driven repayment plan , change your payment due date , or get a deferment or forbearance. NEVER ignore delinquency or default notices from your loan servicer. Consider Simplifying Repayment with Consolidation You might wish to combine your multiple federal student loans into a single Direct Consolidation Loan to simplify repayment. *There may be tradeoffs, so understand the advantages and possible disadvantages of consolidation before you apply.*
[Generated question]: What are the advantages and disadvantages of consolidation?

Table 11: Examples of (a) diverse question generation and (b) unanswerable question filtering. The question generator of ConvRGX generates the questions taking the text in *italics* as grounding span.

ConvRGX Dialogue Generation
[Passage]: In addition to meeting our definition of disability , you must have worked long enough and recently enough under Social Security to qualify for disability benefits. Social Security work credits are based on your total yearly wages or self - employment income. You can earn up to four credits each year. The amount needed for a work credit changes from year to year. In 2019 , for example , you earn one credit for each $1,360 in wages or self - employment income. When you've earned $5,440 , you've earned your four credits for the year. The number of work credits you need to qualify for disability benefits depends on your age when you become disabled. Generally , you need 40 credits , 20 of which were earned in the last 10 years ending with the year you become disabled. However , younger workers may qualify with fewer credits. For more information on whether you qualify , read our publication ; How You Earn Credits. What We Mean By Disability The definition of disability under Social Security is different than other programs.
[Generated Dialogue]: q1 : What are Social Security work credits based on? a1 : Social Security work credits are based on your total yearly wages or self - employment income. You can earn up to four credits each year. q2 : How much do I need for a work credit? a2 : You can earn up to four credits each year. The amount needed for a work credit changes from year to year. In 2019 , for example , you earn one credit for each $1,360 in wages or self - employment income. q3 : What if I earn $5,440 for the year? a3 : you've earned your four credits for the year. q4 : Where can I find more information on whether I qualify? a4 : read our publication ; How You Earn Credits. q5 : Is disability the same as illness? a5 : The definition of disability under Social Security is different than other programs.

Table 12: Example of synthesized dialogue by ConvRGX on the grounding passage.